Visual Exploration of Three-dimensional Gene Expression Using Physical Views and Linked Abstract Views

Gunther H. Weber, *Member, IEEE Computer Society,* Oliver Rübel, Min-Yu Huang, *Student Member, IEEE,* Angela H. DePace, Charless C. Fowlkes, Soile V. E. Keränen, Cris L. Luengo Hendriks, *Member, IEEE,* Hans Hagen, *Member, IEEE,* David W. Knowles, Jitendra Malik, *Fellow, IEEE,* Mark D. Biggin, and Bernd Hamann, *Member, IEEE*

***Abstract*— During animal development, complex patterns of gene expression provide positional information within the embryo and determine cell fates. To better understand the underlying gene regulatory networks, the Berkeley *Drosophila* Transcription Network Project (BDTNP) has developed a suite of methods that support quantitative computational analysis of three-dimensional (3D) gene expression in early *Drosophila* embryos at cellular resolution. We introduce PointCloudXplore, an interactive visualization tool that supports visual exploration of relationships between different genes' expression by combining a variety of established visualization techniques.**

Two aspects of gene expression are of particular interest: (i) gene expression patterns defined by the spatial locations of cells expressing a gene, and (ii) relationships between the expression levels of multiple genes. PointCloudXplore provides users with two corresponding classes of data views: (i) Physical Views provide several different ways to show the spatial relationships between cells and gene expression patterns in the physical embryo, and (ii) Abstract Views discard spatial information and instead plot expression levels of multiple genes with respect to each other, e.g., by 3D scatter plots and bar graphs. In addition, to make more complex analyses possible we use Cell Selectors to highlight data associated with subsets of embryo cells within a View. Further data properties can be revealed by using linking, which can show, in additional views, the data for a group of cells that have first been highlighted by a Cell Selector in an initial view. In this paper we first describe PointCloudXplore as a visualization tool that integrates various established visualization techniques that are useful for exploring 3D gene expression data sets and subsequently provide prototypical examples how it can be used to mine 3D gene expression data sets for new hypotheses.

***Index Terms*— interactive data exploration, three-dimensional gene expression, information visualization, visualization, physical views, multiple linked views, brushing, scatter plots**

G.H. Weber and O. Rübel are with the Computational Research Divisition, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720. E-mail: {GHWeber, ORuebel}@lbl.gov.

M.-Y. Huang and B. Hamann are with the Institute for Data Analysis and Visualization (IDAV), University of California, Davis, One Shields Avenue, Davis. CA 95616, USA. E-mail: {myhuang, bhamann}@ucdavis.edu.

O. Rübel and H. Hagen are with the International Research Training Group "Visualization of Large and Unstructured Data Sets – Applications in Geospatial Planning, Modeling, and Engineering," Technische Universität Kaiserslautern, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany. E-mail: {hagen, ruebel}@informatik.uni-kl.de.

A.H. DePace is with the Department of Molecular and Cellular Biology and the Center for Integrative Genomics, University of California, Berkeley, 142 LSA #3200, Berkeley, CA 94720, USA. E-mail: adepace@berkeley.edu.

C.C. Fowlkes and J. Malik are with the Computer Science Division, University of California, Berkeley, CA 94720, USA. E-mail: {fowlkes, malik}@eecs.berkeley.edu.

S.V.E. Keränen and M.D. Biggin are with the Genomics Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720. E-mail: {SVEKeranen, MDBiggin}@lbl.gov.

C.L. Luengo Hendriks and D.W. Knowles are with the Life Sciences Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720. E-mail: {CLLuengo, DWKnowles}@lbl.gov.

Manuscript received October 13, 2006; revised April 30, 2007.

I. Introduction

THE development of animal embryos is largely controlled by complex networks of transcriptional regulation. In the earliest stages of embryogenesis, a handful of transcription factors are expressed in relatively simple spatial patterns. Over time, these expression patterns become increasingly complex as factors cross-regulate each other and modulate the expression of additional transcription factors in a combinatorial manner. This complex interacting regulatory hierarchy ultimately determines the fate of each cell in the developing zygote [1]. Because each cell separately regulates its own genome, it is essential to determine the locations of cells and quantitate relative levels of multiple genes' expression in them through embryogenesis to accurately describe and computationally model these transcriptional networks.

The BDTNP has chosen the early *Drosophila melanogaster* embryo, the blastoderm embryo, as a model system to explore the formation of gene expression patterns. The basic *Drosophila* body plan is defined during blastoderm stage when the embryo is still morphologically simple. The great wealth of existing knowledge about the regulatory interactions and pattern formation of the *Drosophila* blastoderm make it an ideal model for analyzing genomic regulation of complex patterns. This project has developed a data processing pipeline (Section 2) for extracting precise measurements of spatial patterns of gene expression from imaging data, providing information about the locations of all blastoderm nuclei and the expression levels of a select set of genes associated with each nucleus. However, while *Drosophila* embryos at this stage are already well-researched and existing knowledge can be used to steer the analysis of PointCloud data sets, efficient means to mine these data for new and often unexpected information or properties are missing.

Appropriate visualization has been the key to many major discoveries in science [2]. Visualization uses the immense power and bandwidth of the human visual system to analyze

data, and it is particularly useful for identifying unexpected behavior in a data set. Therefore, we have developed a visualization tool called PointCloudXplore (PCX) that supports visual analysis of 3D gene expression data sets.

PCX is based on two simple, well-established but powerful basic principles. Multiple views (Section IV, Section V) make it possible to show different data aspects without being overwhelmed by the high dimensionality of PointCloud data. Each view emphasizes different data properties, and the interplay between all views makes detailed data analysis possible. The second basic principle is called Cell Selection (called Brushing in Information Visualization) and linking (Section VI). Cell Selection refers to the ability of a user to select data associated with particular groups of cells in any View. Selected data parts (i.e., the Cell Selections) are then highlighted in all data displays. In this way, all views are linked together, making it possible to identify visually which parts of the data in two different views correspond to which group of selected cells.

The multiple views available in PCX can be divided into two groups: (i) *Physical Views* and (ii) *Abstract Views*. Physical Views (Section IV) use information about the volume and position of cells (defined here as a nucleus plus the surrounding cytoplasm) to create different representations of a physical embryo. They show the spatial relationships between cells and expression patterns. Gene expression values are visualized in these views either by *Color Intensity* or by *Expression Surfaces* (surface plots like available in MATLAB and other data analysis tools). Color supports qualitative analysis of gene expression values and the identification of spatial expression patterns. It can be used on either two- or three-dimensional Physical Views. Expression Surfaces provide a more quantitative representation of gene expression data using dedicated surface height plots defined over two-dimensional representations of the embryo.

Abstract views (Section V) use a variety of ways to show the quantitative relationships between multiple genes' expression in one or all cells of the embryo without showing the spatial relationships between cells. We describe two Abstract Views: *2D/3D Scatter Plots* and the *Cell Magnifier* (a 2D histogram plot). 2D/3D Scatter Plots provide a global overview of different genes' expression levels as a function of each other. The Cell Magnifier allows a user to concentrate on one cell and shows information about expression levels in that particular cell. A description of a third Abstract View, *Parallel Coordinates*, is provided elsewhere [3]. While all these techniques are well-established in visualization and widely available in programs such as MATLAB, we describe an integrated system that uses a powerful combination of these techniques and apply and adapt them to 3D gene expression data.

II. Background: Gene Expression and Data Visualization Pipeline

A *Single PointCloud* file contains information about the x, y, z location of each nucleus in an embryo, the nuclear and cellular volumes, and the relative concentrations of gene products (mRNA or protein) associated with each nucleus and cell [4]–[6]. These files are created in the following manner: Embryos are first labeled with two fluorophores to detect two gene products and then with one more to detect the nuclei. Embryos are then mounted, and imaged using a confocal microscope. The obtained images are processed to detect all blastoderm nuclei and measure the fluorescent intensities associated with each gene product in the nucleus and in apical and basal parts of the nearby cytoplasm. For simplicity, in the remainder of this paper we generally refer to the measured fluorescent intensities as gene expression levels, assuming that the two are closely correlated. See Luengo Hendriks et al. [5] for further discussion.

It is not practical to obtain the expression of more than a few genes in a single embryo, due to the limited number of different fluorophores we can distinguish by the microscope as well as the difficulty in adding multiple labels to embryos. Because it is critical to compare the relationships between transcription factors and their many target genes in a common co-ordinate framework, a set of Single PointClouds is registered into one of more *Virtual PointClouds* using both morphology and a common reference gene to determine correspondences [4]. A Virtual PointCloud contains averaged expression levels for many genes mapped onto the nuclei of one of the embryos in the set. PointCloudXplore is used for visualization of both Single PointClouds and Virtual PointClouds.

III. Previous Work

Linking multiple views for the visualization of high-dimensional data sets is an established concept in information visualization [7]. For example, Henze [8] proposed a system for exploring time-varying computational fluid dynamics (CFD) data sets that uses multiple views (called Portraits in his paper) displaying a data set and various derived quantities. Users can perform advanced queries by selecting data subsets in these portraits. The concept of multiple views was also used in the WEAVE system, where a combination of Physical Views and Information Visualization Views (the equivalent of our Abstract Views) allows exploration of cardiac simulation and measurement data [9]. Both Henze's system for CFD data and the WEAVE system use linked views to define features in a data set by refining queries based on brushes. Brushes are equivalent to our Cell Selectors, being highlighted subsets of the data. Doleisch et al. formalized the concept of defining features via queries using Information Visualization Views and utilizing logical operations to combine several brushes [10]. Piringer et al. [11] and Kosara et al. [12] introduced a variety of enhancements to 3D scatter plots, improving depth perception and perception of the sample distribution in all dimensions. Our visualization tool was also inspired by GeneBox [13], which uses scatter plots to visualize results of microarray experiments.

IV. Physical Views: Visualizing Spatial Relationships Between Gene Expression Patterns.

Overview

Physical Views use a 3D embryo model, or different 2D projections of this 3D model, to convey a sense of the spatial

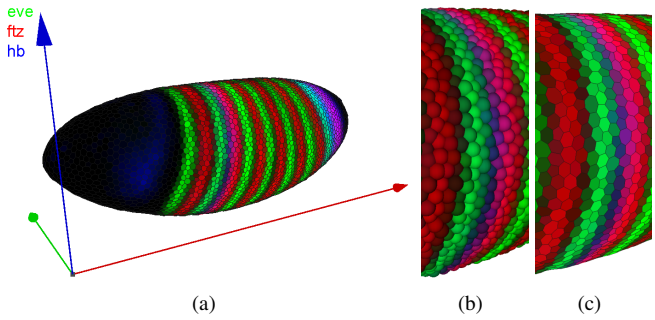


Fig. 1. 3D View (a). Each cell of the embryo is represented by an icon, either a sphere (b) or a polygon in a Voronoi-like tessellation of the surface (c).

distribution of gene expression on the blastoderm. There are three Physical Views in PCX: (i) *3D View*, (ii) *Orthographic View*, and (iii) *Unrolled View*. Each view has its strengths and weaknesses in presenting aspects of gene expression patterns. The 3D View provides the most spatially accurate representation of the embryo; the Orthographic View simulates the 2D views of embryos that most biologists are used to; and the Unrolled View allows expression in all blastoderm cells to be most clearly seen at once, even though the 2D cylindrical projection it employs distorts spatial relationships somewhat. All three views use color intensity to show expression levels similar to the way staining was used to reveal gene expression levels in the original embryo. The 2D views in addition allow graphical Expression Surfaces to be used to portray relative gene expression levels.

3D View

The 3D View utilizes a 3D model of the embryo, which a user can rotate, pan, and scale to obtain an overview of the entire embryo (Fig 1(a)). Figs. 1(b) and 1(c) show that cells can be represented in two ways in this View: as *Spherical Cells* or *Polygon Cells*. In the Polygon Cells View, the embryo is represented by a surface composed of polygonal faces, each of which corresponds to a detected nucleus. These polygons form an approximate Voronoi tessellation of the blastoderm surface and have a visual appearance similar to that of cells. For Polygon Cells, the blastoderm surface is assumed to be a two-manifold surface (being locally flat) and Polygon size depends on the distribution of cells on the embryo blastoderm. Using spherical icons, cells can be shown in embryos with more complex morphological structures that do not form a two-manifold surface. The size of each sphere is chosen relative to the nuclear volume of the cell it represents.

Orthographic View

While the 3D Embryo View provides a user with an intuitive way to manipulate a View of the embryo, a more "traditional" view of the embryo may be desirable, too. A common way to study expression patterns is to use photomicrographic images from defined views, e.g., ventral, dorsal, or lateral view, of the embryo.

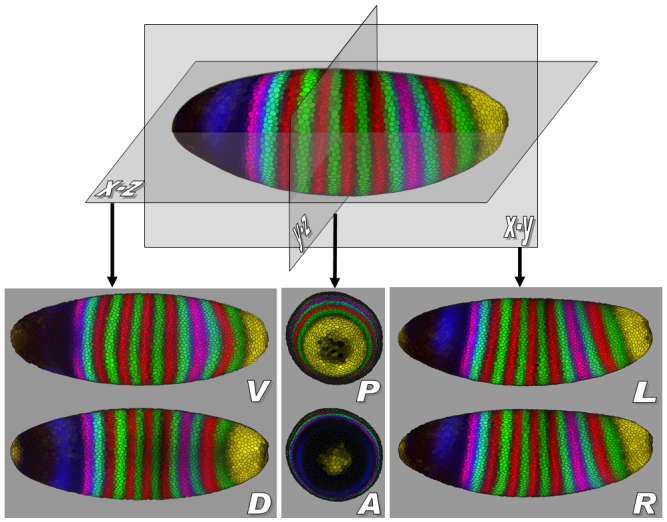


Fig. 2. Orthographic projection views show an embryo along its body axes. The projections shown are of the ventral (V), dorsal (D), anterior (A), posterior (P), left (L), and right (R) views of the embryo.

To simulate these views, we first identify the anterior/posterior (A/P)-axis of the embryo as the smallest eigenvector of the inertia tensor of all cell locations. This eigenvector is equivalent to the axis with the smallest moment of inertia and determined by eigendecomposition of the inertia tensor. We currently determine an embryo's dorsoventral (D/V) orientation manually, based on known expression patterns, and store this information as meta-data in each PointCloud file. These parameters can then be used to rotate the embryo into a standard pose.

Once the embryo is represented in a standard orientation, we project it orthographically along its body axes. To allow a user to gain a global overview, this view is "split". For example, if the left/right (L/R) axis is chosen, the embryo is projected in both directions and the two resulting images, an outside and an inside view of the sides, are shown side by side. This process can be performed analogously for each body axis, see Fig 2. Our tool displays one of these three possible projections at a time: dorsal/ventral (D/V), anterior/posterior (A/P) or left/right (L/R)). As in the 3D View, we use an approximation of the Voronoi tessellation to represent individual cells of the embryo.

Unrolled View

While orthographic 2D projections along body axes provide an overview of the entire embryo, the resulting views are "split" into two "sub-views" (e.g., the left and right sides of the embryo) making it difficult to examine patterns that reach from one side of the embryo to the other. In addition, information at the edges of orthographic projections is compressed. To remedy these shortcomings somewhat, we provide an Unrolled View that maps the entire *Drosophila* embryo continuously to a plane using cylindrical projection [3], [5], see Fig. 3. This is possible because prior to gastrulation, nearly all the cells in the *Drosophila* blastoderm lie in an ellipsoidal monolayer surface. This surface can be "unrolled" by the following process: A

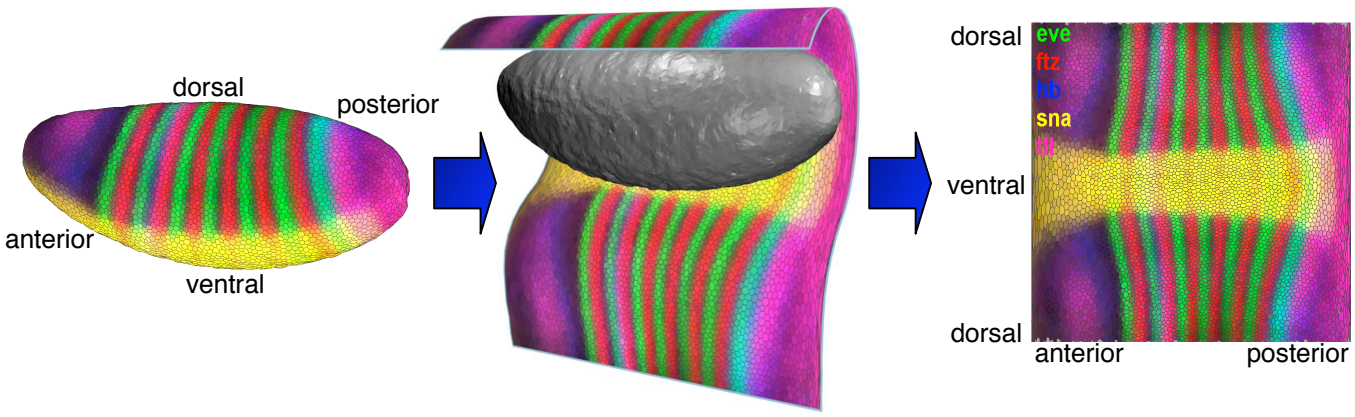


Fig. 3. The Unrolled View uses cylindrical projection to map the entire embryo to a 2D plane.

standard orientation of the embryo is used as in the projection view and the embryo is surrounded with a cylinder whose axis is aligned with the embryo's A/P-axis. All detected cells on the blastoderm surface are then projected onto this cylinder. The resulting surface is then unrolled by cutting the cylinder along a line corresponding to the dorsal midline of the embryo. This process yields a continuous mapping of the embryo surface to a 2D plane and allows users to trace expression patterns over the entire embryo.

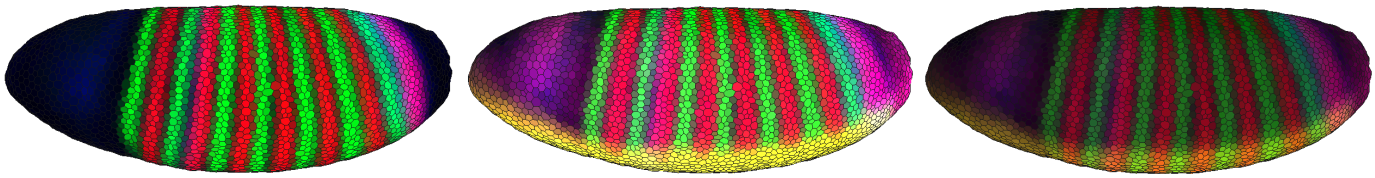
Visualizing Gene Expression via Simulated Staining

Given any of the above graphical representations of the embryo, we need to visualize gene expression levels on that model. In acquired images, stain brightness indicates the relative expression levels of tagged gene products within an embryo. PointCloud data contains corresponding expression information of two or more genes, and this information can be transferred into the graphical representation of the embryo. A user can then select a subset of genes and the color of a simulated "fluorescent stain" for each gene. For color selection, we use the hue-saturation-value (HSV) color model as it represents color by hue (i.e., the actual color such as red, blue, green or yellow), saturation (i.e., a measure that specifies how much the color differs from gray) and a value that specifies color brightness (i.e. intensity). This representation makes it possible to select a set of colors that can be easily distinguished by the eye and also to independently choose the brightness of a color according to the expression level of the corresponding gene. If more than one gene is expressed in a cell, we calculate a color at appropriate brightness for each expressed gene and mix the resulting colors to obtain a color for the cell.

Fig. 4 shows the staining patterns for either three genes or five genes on a 3D View. If no more than three genes are of interest to the user, it is advantageous to choose staining pattern colors from red (with a hue of $0°$), green (with a hue of $120°$) or blue (with a hue of $240°$), the three basic colors of the additive red-green-blue (RGB) color model, which is used to display colors on the screen (Fig. 4(a)). These colors allow each combination of expression levels to map to a unique mixed color.

If expression levels of more than three genes are of interest, it is no longer possible to choose colors that are independent in the RGB color model. Since the colors are no longer independent, a given mixed color can be obtained by more than one combination of gene expression levels, see Figs 4(b) and 4(c). In addition, it becomes possible to have "overexposed" cells. This overexposure arises due to the way colors are mixed in the RGB color model. The intensity of each component is represented by a real value ranging from zero (no contribution of this component) to one (component at full intensity). A cell color is obtained by adding all individual color components for all selected genes. If any given component exceeds an intensity of one, it is clamped, i.e., it is set to one. This behavior is visible on the ventral (lower) side of the embryo shown in Fig 4(b). Here, most cells are yellow (red and green at full intensity) since *snail (sna)* expression is mapped to yellow. Consequently, adding more red or green for *fushi tarazu (ftz)* and *even-skipped (eve)* stripes does not change the mixed color.

To gain a better overview when "overexposure" occurs, we allow the user to specify a global weight for all expression pattern contributions. This weight, ranging from zero to one, is multiplied with all expression level colors before they are mixed. Choosing a smaller weight can be thought of as reducing the exposure time of a photograph. All colors become darker and thus colors need to be clamped less frequently. For example, in Fig. 4(c) a smaller display weight was chosen. The yellow *sna* pattern no longer corresponds to maximum red and green intensity. Thus, the *ftz* and *eve* stripes become visible in the formerly "overexposed" region. Aside from manual definition of the described global weight, PCX also provides an auto exposure function that automatically sets the global weight to an appropriate value. To obtain this automatic weight, mixed colors for all cells are calculated assuming a weight of one. Subsequently, the maximum intensity of all color components of all cells is computed. The automatic weight is then chosen as the reciprocal of that global maximum intensity. This choice has the effect that no cell has a color component exceeding a value of one.



(a) (b) (c)

Fig. 4. Gene expression patterns are mapped onto the iconographic cell representations as color. (a) shows the expression of three genes *ftz* (red), *eve* (green) and *hb* (blue). Here, each combination of expression levels maps to a unique color. In (b) and (c) two additional genes, *sna* (yellow) and *tll* (pink), are shown on the embryo. Mixed colors are no longer unique. In (c) the brightness (i.e., a global weight for each color) is decreased to show variation in the "overexposed" yellow and white regions on the ventral (lower) side of the embryo.

The intensities of expression patterns can be further altered for each gene separately using thresholding of the measured fluorescence intensity values. Thresholding means that minimum and maximum cutoff values are specified. If the expression level within a cell is below the minimum cutoff value, the expression is mapped to the minimum intensity (black). Similarly, if the expression is above the maximum threshold, it is clamped to the maximum intensity value.

For each gene, we provide the user with a histogram that plots the number of cells in which specific fluorescence intensities were measured. These histograms are overlaid with minimum and maximum cutoff values and sliders that allow the user to alter the threshold. When a user changes the maximum and minimum thresholds, information is provided on the percent of cells in the embryo that are below the minimum threshold (i.e., cells that are displayed unstained), the percent of cells that are above the maximum threshold (i.e., cells that are displayed with maximum stain intensity), and the percent of cells that are in the chosen threshold interval (i.e., cells that are mapped to an intermediate stain brightness), aiding a user in the appropriate choice of these thresholds. Colors in the various views are updated immediately while changing the minimum or maximum value of a gene, allowing interactive validation of the effects and appropriateness of the values.

Great care must be taken in using this thresholding strategy. The gene expression data in PointCloud files is measured data and as such is subject to noise. Moreover, since all gene expression is normalized from zero to one hundred, regardless of the actual expression levels, expression patterns with lower intensities are obscured with higher levels of noise. Slight biases in attenuation correction are also likely to influence the detectability, symmetry and shape of the patterns (for further discussion, see Luengo Hendriks et al. [5]). Thresholding can be used to reduce noise by, for example, setting a weak background staining to zero, making the actual gene expression pattern clearer, see Figs 5(a) and 5(b). This strategy is most useful when multiple genes are being displayed as the cumulative effect of several backgrounds can confuse the view. However, it is frequently not clear what part of PointCloud data is noise and what is signal. The user must be aware that key biological information can be obscured by thresholding. With this caveat in mind, thresholding can be helpful in enhancing the view of some "real" features of an expression pattern even if other significant features are at the same time obscured. Extreme thresholding can be used to emphasize certain expression properties. Figs. 5(c) and 5(d), for example, were obtained using a very low maximum threshold to illustrate that the interstripe expression levels of *even-skipped (eve)* are typically higher than *eve* expression levels on the anterior and posterior of the embryo.

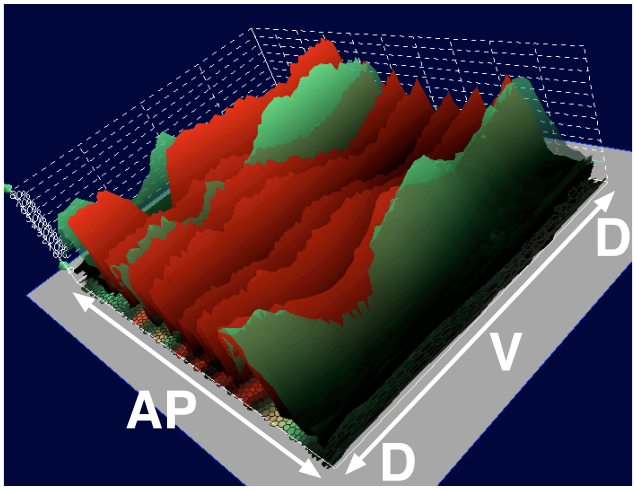


Fig. 6. Gene Expression Surfaces show quantitative gene expression levels. In the figure two surfaces show expression levels for *eve* (red) and *hb* (green). The direction of the anterior/posterior (AP) axis and the dorsal/ventral (DV) circumference are shown.

Gene Expression Surfaces

In addition to providing an overview of the entire embryo, projecting the embryo to a plane has the advantage of freeing one dimension up for displaying additional information. In PCX, this "free" dimension can be used to display gene expression values as offset surfaces, so-called *Gene Expression Surfaces* [3], which allow a more quantitative analysis of gene expression data. Expression Surfaces can be defined over the Orthographic or the Unrolled Views. Each Expression Surface displays data for one gene. The xy-positions of Expression Surface points are determined by the positions of cells in the underlying views, whereas the height of an Expression Surface is determined by the expression values measured for the gene it represents. The height mapping of Expression Surfaces is defined consistently with the color mapping used on the model of the embryo, i.e., the minimum and maximum gene

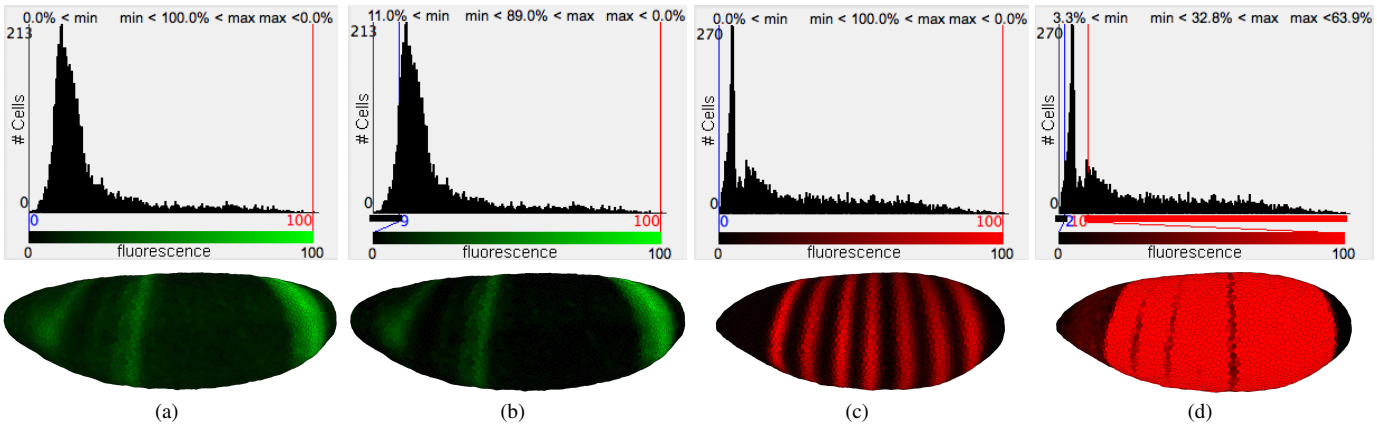


Fig. 5. Using histograms that show the number of cells expressing a gene as a function of the expression levels makes it possible for a user to choose different cutoff values to allow aspects of the data to be better shown. (a) and (b) illustrate this concept by thresholding *hunchback (hb)* to reduce background noise. (c) and (d) use an extremely low maximum threshold for *even-skipped (eve)* (d) to illustrate that the inter-stripe expression levels of *eve* are typically higher than its expression levels on the anterior and posterior of the embryo.

threshold values described above are also applied here. Thus, Expression Surfaces provide an additional way for assessing the use of threshold values. Spatial relationships between several genes' expression patterns can be viewed at once using multiple Expression Surfaces. For example, Fig. 6 shows the quantitative relationship between mRNA expression patterns of the transcription factor *hb* (green) and its target gene *eve* (red), showing that along the dorsal ventral axis, the ratio of the two genes' expression levels change. Thus, they may not have as simple a quantitative relationship as some current models assume.

A variety of options are provided to improve the view, including different coloring strategies and changing Expression Surface transparency or intensity values. An Expression Surface can, for example, be of the same color as the gene in the 2D plot, but have an intensity that varies with gene expression levels to allow comparison of expression patterns using both color and height in parallel.

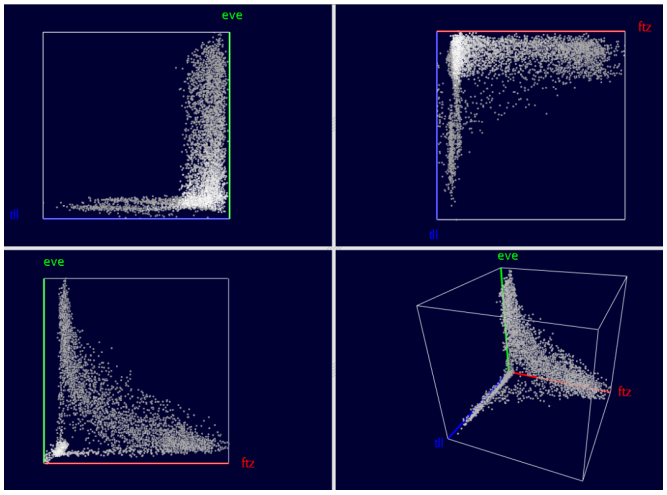


Fig. 7. Scatter Plots show the relationship between up to three genes' expression levels.

V. Abstract Views

To understand regulatory dependencies within a network, it is important to compare expression levels of transcriptional regulators to those of their target genes. We believe that important regulatory relationships may well be discernible from comparisons of relative expression levels in cells regardless of the relative locations of cells to one another. To detect possible relationships based only on gene expression levels, we use information visualization techniques that reveal these relationships in *gene expression space*. Different Abstract Views each explore gene expression space by a variety of means. We describe two Abstract Views: Scatter Plots and the Cell Magnifier.

2D/3D Scatter Plots

Scatter Plots, see Fig. 7, are the conceptually simplest way to visualize relationships in gene expression space. In this View, three genes are selected and mapped to the three axes of a Cartesian coordinate system. Each coordinate axis represents one gene's expression level ranging from no expression at the origin to maximum relative expression. Each cell in the embryo is represented by a single point in the 3D scatter plot with the point location specified by the relative gene expression levels.

To better distinguish separate points and to estimate their location, we use colors, halos, and alpha blending [11], [12]. To enhance the depth perception of points we decrease their brightness as their distance to the viewer increases. Close points are drawn with a brighter color than distant points, and show up more prominently. (We only adjust point color brightness since the actual point color is used to convey additional information during Cell Selection, see Section VI.)

In dense regions of a Scatter Plot it is also difficult to distinguish individual points. "Halos" make it possible to better distinguish points in these dense regions. Drawing a disc around each point (similar to the corona around the sun

during a solar eclipse) keeps points of similar color in dense regions of a scatter plot distinguishable. Another method of dealing with high point densities in scatter plots is the use of transparency. Points are plotted partially transparent, i.e., points in the background "shine through" points on the front.

Fig. 7 shows the basic layout of our 3D Scatter Plots. The 3D Scatter Plot (lower-right panel) that shows the relationship between three genes is augmented with a set of 2D Scatter Plots that show expression relationships between the three possible pairs of genes in the 3D plot. We chose this layout as the 2D plots always provide a "standard" view on the data, while the viewpoint for the 3D Scatter Plot can be chosen arbitrarily by rotating the plot. (Furthermore, the 2D Scatter Plots facilitate Cell Selection, see Section VI.) Looking at Scatter Plots of expression levels alone can reveal information about relationships in gene expression. For example, the Scatter Plots shown in Fig. 7 exhibit the expression relationship between *ftz*, *eve* and *tailless (tll)*. The lower-left panel shows the anti-correlation between *ftz* and *eve* that express in alternating stripes in most cells. However, the Scatter Plot in that panel also shows that cells exist that express neither *ftz* nor *eve* strongly. By consulting the 2D Scatter Plots corresponding to the other gene combinations *eve/tll* (upper-left panel) and *ftz/tll* (upper-right panel), it becomes obvious that *tll* expression is high in cells that do not express *ftz* or *eve*. This fact becomes also apparent when one considers the 3D Scatter Plot that shows the relationship between all three genes. Here, it is immediately obvious that cells with high *tll* levels express neither *ftz* nor *eve* strongly, which is consistent with the fact that high *tll* expression occurs at the anterior and posterior ends of the embryo where *ftz* and *eve* stripes end, see Figs. 1 and 4.

Cell Magnifier

Unlike the other Physical and Abstract Views currently available in PCX, the Cell Magnifier (Fig. 8, right panel) concentrates not on comparing gene expression values in different cells but on comparing expression values in just one cell. In this View, it is not the expression values in other cells that provide the surrounding context but all expression values measured in the current cell itself. In the Cell Magnifier, gene expression values are visualized using a bar graph, one bar for each gene. The individual bars are colored according to the user defined stain colors. Since exact expression values can only be roughly estimated from bar size, the exact measured gene expression value is also displayed beside each bar. The cell to be displayed in the Cell Magnifier can be selected in any Physical View and is highlighted by graying it out (arrowed in Fig. 8, left panel). The Cell Magnifier in the right panel of Fig. 8 shows the gene expression profile of a cell in the most anterior *eve* stripe.

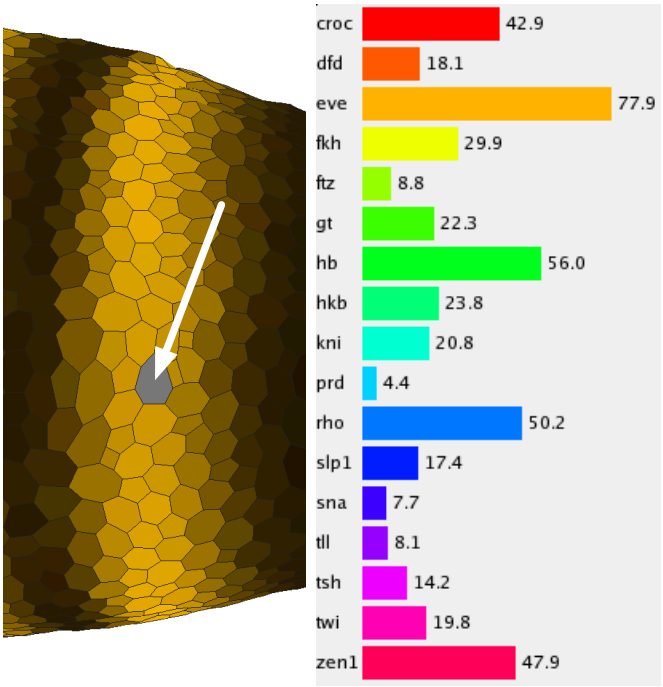


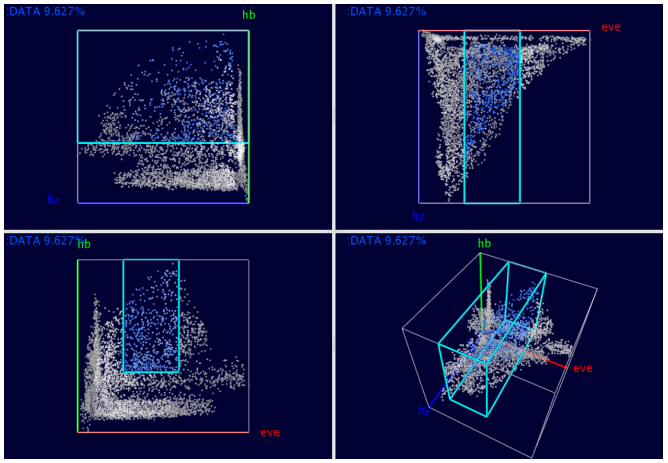
Fig. 8. The cell magnifier supports comparison of multiple genes' expression levels within a single cell.

VI. Cell Selection and Linking

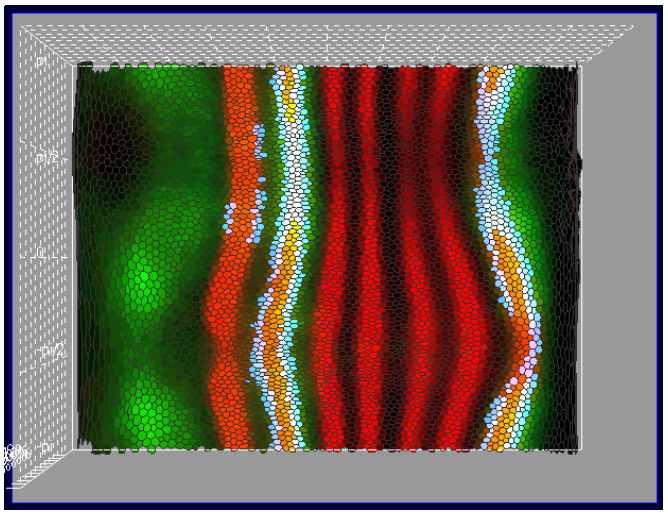
All of the Physical and Abstract Views that we have described are useful in their own right and can be used individually to mine data sets for new information. However, it is often desirable to correlate information shown in different views. For example, when looking at a Scatter Plot of *hunchback (hb)*, *eve*, and *ftz* one might be interested in concentrating on cells with high *hb* and medium *eve* expression and ask where on the embryo these cells are located.

Cell Selection and linking provide effective means to perform queries of this type. Cell Selection is the selection of a subset of cells. The selection is subsequently highlighted in a view using color (chosen in the same way as stain colors by specifying a hue in the HSV color model). Cell Selection can be performed in any View. Fig. 9 illustrates this concept using a Scatter Plot for the genes *hb*, *eve*, and *ftz*. A user can select a rectangular box in the 3D Scatter Plot. This box defines a minimum and maximum threshold for each of the three displayed genes. Since moving a box in a 3D scatter plot with a 2D input device such as a mouse can be difficult, PCX also shows the projection of this box as a rectangle in each 2D Scatter Plot. In these 2D plots, it is easier to move or resize the region of selected cells.

In Fig. 9(a) cells with high *hb* and medium *eve* expression are highlighted and colored in light blue. By linking the selection in the Scatter Plot to a Physical View, such as the Unrolled View in Fig. 9(b), a user can identify which cells in the embryo express genes at the selected levels and consequently relate expression level relationships to physical patterns of cells in the embryo. All views are updated simultaneously during the selection process, allowing a user to follow the changes of the pattern of selected cells during the selection process. We note that showing selected cells in a Physical View adds a new simulated stain corresponding to a binary expression pattern. The final color of a cell is obtained by mixing this "selection stain" with all other "active stains". Because of this mixing, it is possible that selected cells have different colors in the



(a)



(b)

Fig. 9. Cell Selection and Linking allow a user to relate different views to each other. For example, it is possible to select a subset of cells based on expression levels of three genes, such as *hb*, *ftz* and *eve*, in a Scatter Plot view (a) while highlighting the same cells in a Physical View, such as the Unrolled View (b) and show their spatial distribution.

Scatter Plot View and a linked Physical View.

Linking can connect any types of views, including views of the same type. Fig. 10 shows a second Scatter Plot View that is linked to the Scatter Plot shown in Fig. 9(a). Here, a different set of genes, *hkb*, *hb*, and *kni*, is mapped to the three coordinate axes. By selecting cells based on the expression levels of *hb* and *eve* and showing the expression of the same cells with respect to *hkb*, *hb*, and *kni*, it is possible to explore expression relationships between more than three genes. For example, notice how Fig. 10 shows that the selected cells with high *hb* and medium *eve* expression all show low expression of *hkb* and *kni*. Cell Selection is possible in Physical and Abstract Views. In Physical Views, selection is performed by "painting" patterns on the embryo, see, for example, Fig 11. In Fig. 11(a), the three *hb* stripes have been assigned to three different Cell Selectors shown in red (anterior stripe), blue (center stripe) and light green (posterior stripe) in a 3D View. By linking a Scatter Plot View (Fig. 11(b))

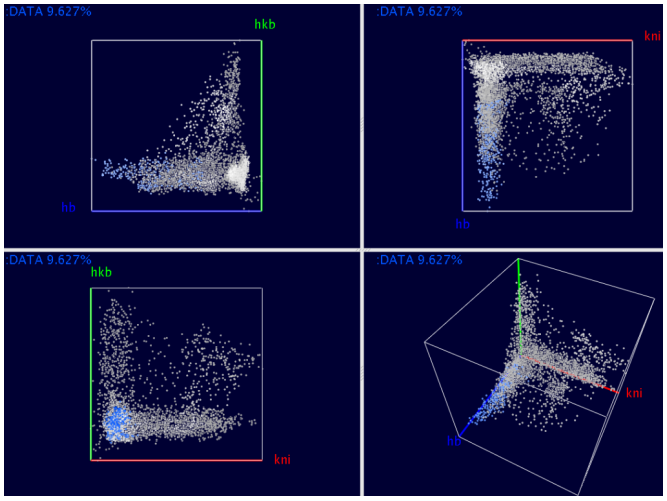
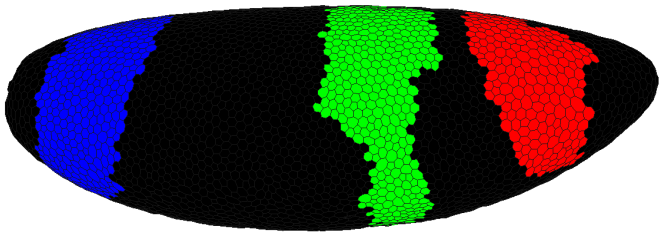
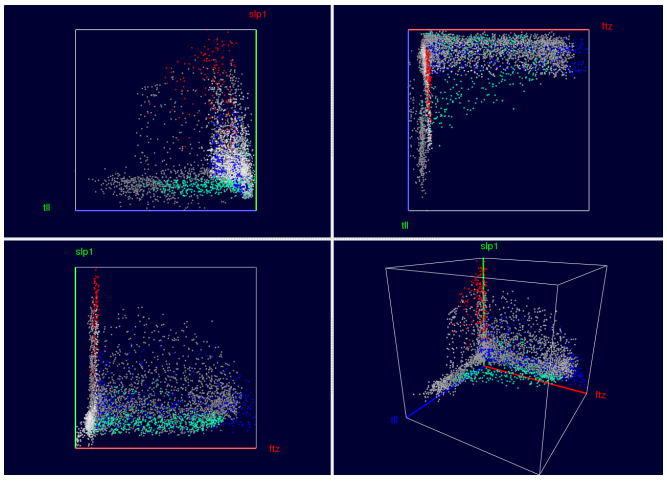


Fig. 10. PCX supports Linking of arbitrary types of views. For example, it is possible to link two Scatter Plots (for example, the Scatter Plot shown in Fig. 9(a) and the Scatter Plot shown in this figure) of different gene combinations.



(a)



(b)

Fig. 11. Cell Selection in the embryo view. In the embryo view it is possible to "paint" the three *hb* stripes, assign them to three distinct Cell Selectors and show them in three distinct colors. By linking these selections to a Scatter Plot view (here showing *ftz*, *slp1* and *tll*) a user can determine how these genes are expressed differently within the individual stripes.

to the 3D View it becomes possible to emphasize different expression behavior within the stripes. The Scatter Plot in the right panel shows this difference for the genes *ftz*, *slp1*, and *tll*. For example, it becomes apparent that the anterior stripe (red) has generally a higher expression of *slp1* than the other two stripes. Furthermore, when considering this three-gene combination, the three peaks of *hb* RNA expression seem to form three clusters in the Scatter Plot.

Painting cells manually on the embryo can be rather time-consuming. If a user is interested in defining a contiguous region of cells on the embryo with similar expression behavior, it is possible to first color the model of an embryo in a Physical View with selected genes and then paint regions using the gene expression colors as a guide. If higher accuracy is desired, one can examine cells by means of the cell magnifier and add only those cells with expression levels in a certain range. However, to automate this process, PCX provides *Seed Cell Selection*, which employs a cell selected using the Cell Magnifier, such as the gray cell shown in the left panel of Fig. 12. The user then selects one or more genes whose expression level(s) should be considered in defining the region of the embryo, such as *ftz* in Fig 12 (center panel). Seed cell selection then uses a flood fill method [14] to identify all cells in a contiguous region whose expression levels lie within the specified expression range(s), see Fig. 12 (right panel).

VII. Combining Cell Selectors into Complex Queries

All views available in PCX are linked via a central *Cell Selector Management System*. Cell Selectors of any kind can be accessed here in a unified way and common Cell Selector properties, such as color, can be defined. Furthermore, the central Cell Selector management window allows one not only to perform basic management operations, but also supports combining Cell Selectors using logic operations, such as AND, OR, and NOT. Thus, for example, Cell Selectors defined in a Scatter Plot (DATA Cell Selectors) can be combined with Cell Selectors defined by drawing on the embryo or via seed cell selection (POSITION Cell Selectors), making it possible to define higher-order cell selections within gene expression space.

Logical Cell Selectors are a subset of Cell Selectors that are not manually defined by the user, but are computed by a logical operator using other Cell Selectors as input. The NOT operator, for example, inverts the selection defined by another Cell Selector, i.e., it selects all cells which are not selected by the given Cell Selector. The AND and the OR operations take the selection defined by two Cell Selectors as input. OR combines both selections by computing the union of two Cell Selectors, i.e., it selects all cells that are selected by the first or the second Cell Selector. The AND operator computes the intersection of two Cell Selectors, i.e., it selects only those cells that are selected by both the first and the second cell selector. Since logical operators create new (Logical) Cell Selectors it is not only possible to display results of Cell Selector combinations in any view but also to use Logical Cell Selectors as inputs to other logical operations and form complex queries.

Fig. 13 shows an example of Cell Selector combinations. In general, the genes *giant (gt)*, *hb*, and *Krüppel (Kr)* are accepted as regulators of the second stripe of the *eve* expression pattern from the anterior to the posterior of the embryo. To illustrate this fact using PCX, we first classify the expression patterns of these three genes by creating an individual Cell Selector for each gene by defining a range of gene expression using a Scatter Plot and linked Unrolled View. For example, in Fig. 14 the Scatter Plot shown in the right part of the window is used to define a threshold for *gt* while the selection is validated interactively by comparing the spatial pattern defined by the selection with the *gt* expression pattern using an expression surface. Genes *Kr* and *gt* are both known to be repressors while *hb* is an activator of *eve* stripe two [15]. Therefore, we invert the expression patterns of *Kr* and *gt* using NOT operations. Afterwards, the *hb*-Cell Selector and the Logical Cell Selectors that define the inverted patterns of *Kr* and *gt* are combined using a sequence of AND operations to compute the intersection of these three patterns (colored green in Fig. 13). The spatial pattern resulting from this selection is then compared to the *eve* expression pattern (colored red in Fig. 13), which has also been classified by a Cell Selector via manual thresholding in a Scatter Plot (Fig. 13, left). The resulting overlay shows that the second eve expression stripe coincides well with a defined stripe-like region (yellow region in Fig. 13) formed by the complex Logical Cell Selector described above, consistent with the view that *hb*, *Kr*, and *gt* are responsible for the formation of *eve* stripe two. For *eve* stripe seven only the anterior border of the stripe follows the border of another stripe-like region formed by the described selection. This observation may be interpreted as an indication that *hb*, *Kr*, and/or *gt* are also involved in regulation of eve stripe 7 but that additional regulatory factors are needed for a complete definition of this stripe. However, it should be remembered that coexpression does not necessarily imply positive regulatory interaction or anticorrelation a negative regulatory interaction—for example, high levels of *hb* are actually known to repress stripe seven. As shown in Fig. 13, it is possible to define and represent complex queries by a simple tree structure. Cell Selectors defined by the user always appear as leaf nodes of such a tree since they do not rely on the input of other Cell Selectors. Further, Logical Cell Selectors are always inner nodes of a Cell Selector tree. As illustrated in the above example, visual validation of selection results is facilitated when first treating different genes independently before defining more complex cell queries.

VIII. User Interface

An important consideration during the development of PCX was to keep the graphical user interface (GUI) as simple as possible without limiting its power. We incorporated many rounds of feedback from the biologists who are end-users of the system in order to provide fast and easy access to all views and system controls. Fig. 14 shows a snapshot of the GUI of PCX. The main window is split into two main areas: The left part contains all Physical Views of the embryo (for example, Expression Surfaces over an Unrolled View in Fig. 14) and the

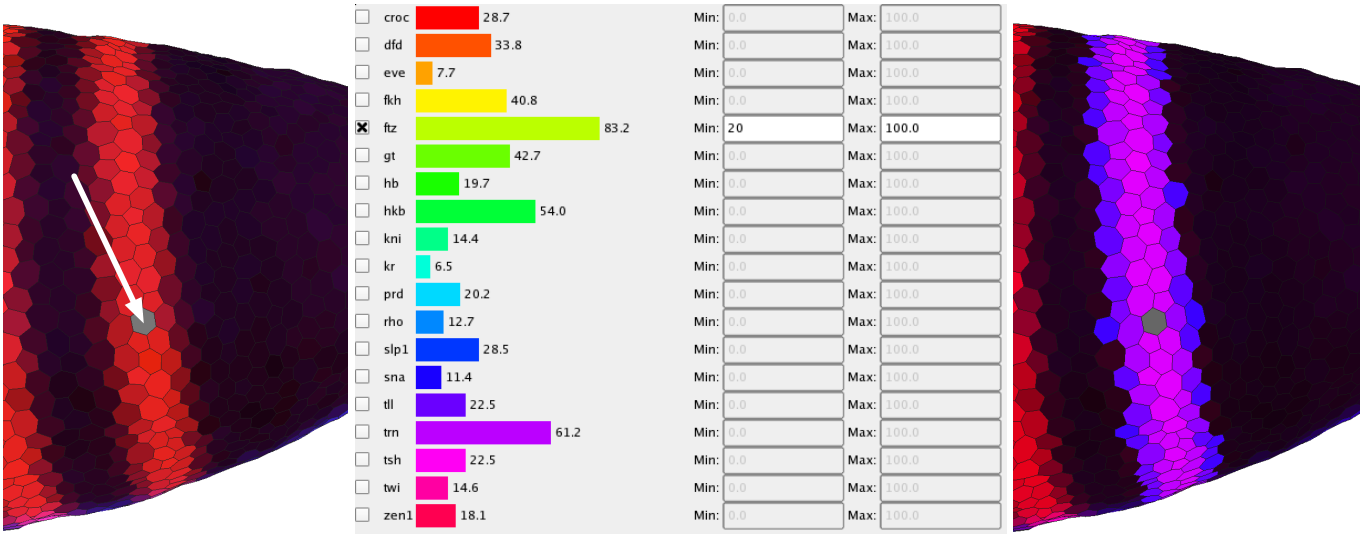


Fig. 12. Using seed Cell Selection to select all cells belonging to a *ftz* stripe.

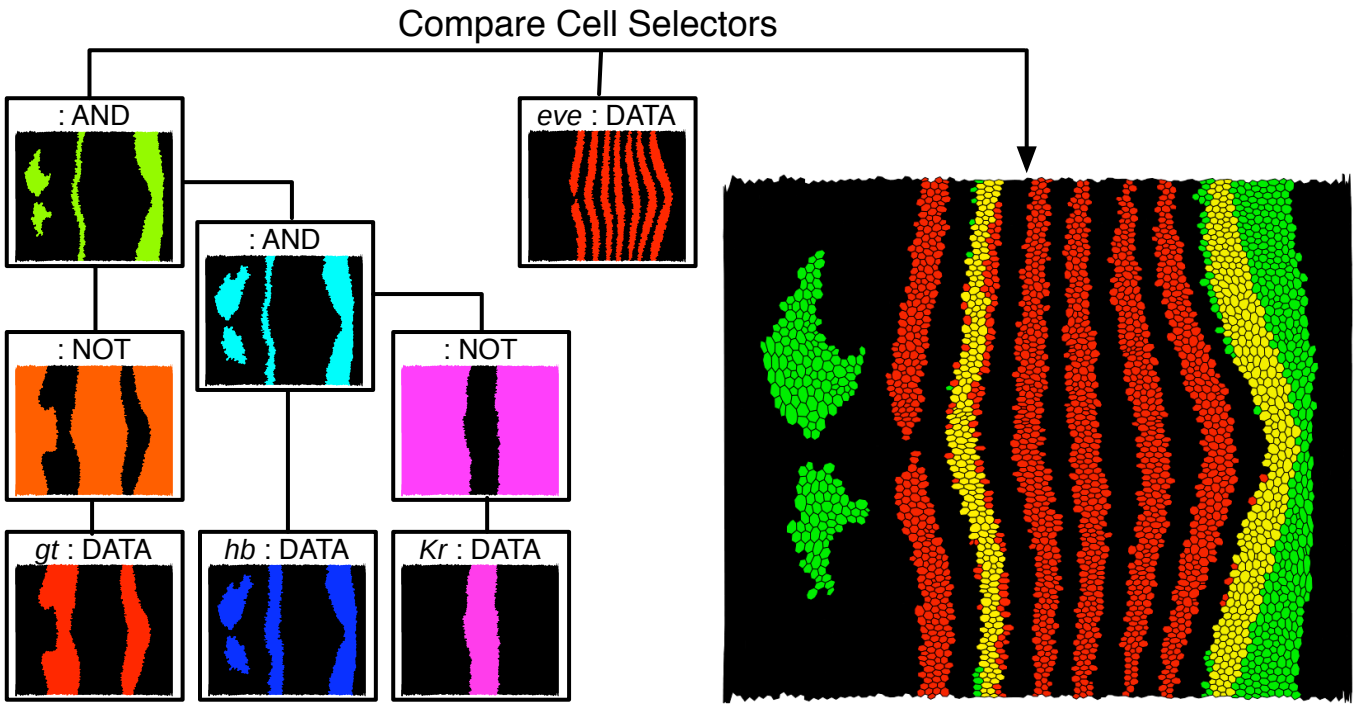


Fig. 13. The expression patterns of *giant (gt)*, *hunchback (hb)*, and *Krüppel (Kr)* are first classified by defining an independent Cell Selector in Scatter Plots (DATA cell selector) for each gene. Subsequently, the Cell Selectors defining the *gt* and *Kr* patterns are inverted using a NOT operation. Afterwards these Logical Cell Selectors as well the Cell Selector defining the *hb* pattern are combined using a sequence of AND operations. In this way the overlap of the hb expression pattern, and the inverted *gt* and *Kr* expression patterns can be determined. The result (green) is compared to the *eve* expression pattern (red) identified by another DATA Cell Selector.

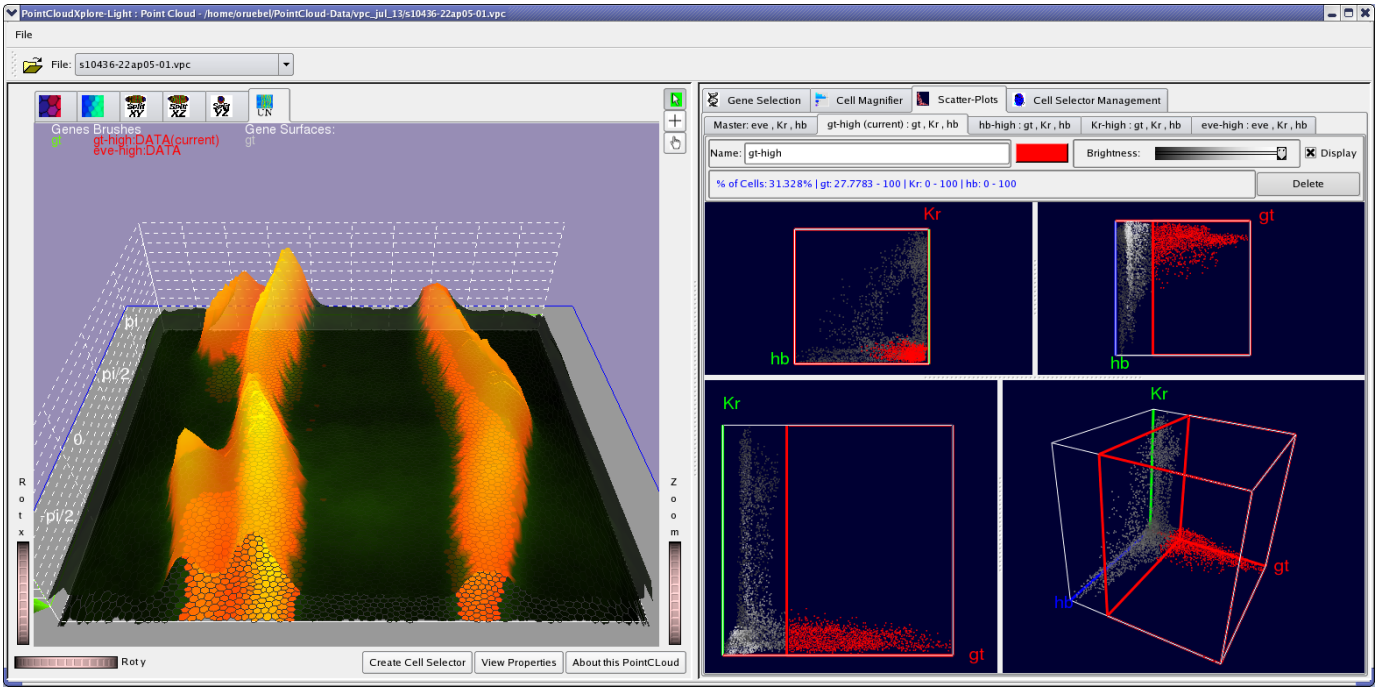


Fig. 14. The GUI of PCX. The window is split into two major parts. The left part contains all Physical Views while all other views, as well as additional user controls, are arranged in detachable tabs in the right part of the window. In the Scatter Plot tab one can also create additional sub-tabs each containing one scatter plot. Each of these additional Scatter Plots is responsible for editing one specific Cell Selector.

right part all Abstract Views as well as additional user controls (for example, a Scatter Plot View in Fig. 14). Controls and Abstract Views in the right part are arranged in a series of detachable tabs such that a user can switch between different abstract views or detach them from the main window and show them side-by-side. The Scatter Plot View can hold several sub-tabs corresponding to active Cell Selectors. It always holds a so-called "Master" Scatter Plot tab (leftmost tab in Scatter Plot View), which is used to choose gene combinations for creating new Cell Selectors. Each newly created Cell Selector shows up as a separate tab. For example, Fig. 14 shows four tabs for cell selectors named "gt-high", "hb-high", "Kr-high", and "eve-high". This strategy has the advantage that it is simple to switch between editing different Cell Selectors (and associated Scatter Plots). To change a Cell Selector, the user simply selects the corresponding tab and does not have to worry about selecting the correct combination of genes in order to edit a cell selector.

IX. Results

- Examples showing the value of PointCloudXplore.

X. Implementation

PointCloudXplore is an interactive exploration tool. Views of the embryo are rendered interactively and all interactions that we described take at most the fraction of a second to complete. We implemented PointCloudXplore as C++ stand-alone application using Trolltech's Qt 4.2 library (`http://www.trolltech.com/`) and OpenGL (`http://www.opengl.org/`) as cross-platform widget and graphics libraries. We have compiled and run PointCloudXplore successfully on Linux, MacOS X and Windows machines. PointCloudXplore is currently available from the BDTNP's web page `http://bdtnp.lbl.gov/Fly-Net/bioimaging.jsp?w=pcx` as a means to view the project's release data set. At the time of this writing, only single PointClouds describing gene expression of a single embryo are publicly available, limiting the public use of our tool as these data sets are noisy and only comprise of expression information of two genes. The project intends to release PointClouds describing gene expression of n (30?) genes in the near future.

XI. Extensions

Scatter Plots are limited in so far that only three genes can be displayed at once. While it is possible to show expression relationships between more than three genes by linking two or more scatter plots, there are other methods for visualizing high-dimensional expression space. Parallel Coordinates are commonly used for that purpose and we have integrated them into PCX. Further details can be found in another paper [3].

XII. Conclusions and Future Work

The combination of several views for visualizing 3D gene expression information has proven to be a valuable tool to members of the BDTNP in finding new relationships in 3D gene expression data. We plan to release a simplified version of this tool as part of the first data release of the project, allowing biologists from other groups to explore published PointCloud data. As the BDTNP collects PointCloud data for many more

genes, several additional challenges will arise that we plan to address. One of these challenges is mapping an even larger number of genes to colors. With our current approach, we can show expression levels of five to six genes simultaneously without confusing a user. (The exact number depends on the spatial distribution of the patterns.)

Scatter Plots and Parallel Coordinates should allow the relationship between 20-30 genes to be examined. In the future however, we anticipate the need to examine several hundred genes at once in Virtual PointClouds. We hope to address this challenge by combining our visualization tool with automated data analysis methods, such as clustering or self-organizing maps to reduce dimensionality of data sets and define new methods of mapping gene expression level combinations to colors.

ACKNOWLEDGMENT

This work was supported by the National Institutes of Health through grant GM70444, by the National Science Foundation through award ACI 9624034 (CAREER Award), through the Large Scientific and Software Data Set Visualization (LSSDSV) program under contract ACI 9982251, and a large Information Technology Research (ITR) grant; and by the LBNL Laboratory Directed Research Development (LDRD) program; A.H. DePace is funded by a Helen Hay Whitney postdoctoral fellowship. We thank the members of the Visualization and Computer Graphics Research Group at the Institute for Data Analysis and Visualization (IDAV) at the University of California, Davis; the members of the BDTNP at the Lawrence Berkeley National Laboratory (LBNL) and the members of the Visualization Group at LBNL.

REFERENCES

[1] P. A. Lawrence, *The Making of a Fly: The Genetics of Animal Design*. Blackwell Publishing, Inc., 1992.

[2] E. R. Tufte, *The Visual Display of Quantitative Information*. Graphics Press, 1992.

[3] O. Rübel, G. H. Weber, S. V. E. Keränen, C. C. Fowlkes, C. L. Luengo Hendriks, L. Simirenko, N. Y. Shah, M. B. Eisen, M. D. Biggin, H. Hagen, D. W. Knowles, J. Malik, D. Sudar, and B. Hamann, "Pointcloudxplore: Visual analysis of 3d gene expression data using physical views and parallel coordinates," in *Data Visualization 2006 (Proceedings of Eurographics/IEEE-VGTC Symposium on Visualization 2006)*, B. Santos, T. Ertl, and K. Joy, Eds. Aire-la-Ville, Switzerland: Eurographics Association, May 2006, pp. 203–210.

[4] C. C. Fowlkes, C. L. Luengo Hendriks, S. V. E. Keränen, M. D. Biggin, D. W. Knowles, D. Sudar, and J. Malik, "Registering *Drosophila* embryos at cellular resolution to build a quantitative 3d map of gene expression patterns and morphology," in *CSB 2005 Workshop on BioImage Data Minning and Informatics*, August 2005.

[5] C. L. Luengo Hendriks, S. V. E. Keränen, C. C. Fowlkes, L. Simirenko, G. H. Weber, A. H. DePace, C. Henriquez, D. W. Kaszuba, B. Hamann, M. B. Eisen, J. Malik, D. Sudar, M. D. Biggin, and D. W. Knowles, "Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution I: data acquisition pipeline," *Genome Biology*, vol. 7R124, 2006, doi:10.1186/gb-2006-7-12-r124.

[6] S. Keränen, C. Fowlkes, C. Luengo Hendriks, D. Sudar, D. Knowles, J. Malik, and M. Biggin, "Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution II: dynamics," *Genome Biology 2006*, vol. 7:R124, 2006, doi:10.1186/gb-2006-7-12-r124.

[7] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky, "Guidelines for using multiple views in information visualization," in *AVI '00: Proceedings of the working conference on Advanced visual interfaces*. New York, NY, USA: ACM Press, 2000, pp. 110–119.

[8] C. Henze, "Feature detection in linked derived spaces," in *Proceedings IEEE Visualization '98*, D. Ebert, H. Rushmeier, and H. Hagen, Eds. Los Alamitos, CA, USA: IEEE Computer Society Press, 1998, pp. 87–94.

[9] D. L. Gresh, B. E. Rogowitz, R. L. Winslow, D. F. Scollan, and C. K. Yung, "WEAVE: A system for visually linking 3-d and statistical visualizations, applied to cardiac simulation and measurement data," in *Proceedings IEEE Visualization 2000*, T. Ertl, B. Hamann, and A. Varshney, Eds. Los Alamitos, CA, USA: IEEE Computer Society Press, 2000, pp. 489–492.

[10] H. Doleisch, M. Gasser, and H. Hauser, "Interactive feature specification for focus+context visualization of complex simulation data," in *Data Visualization 2003 (Proceedings of the Eurographics/IEEE TCVG Symposim on Visualization)*, G.-P. Bonneau, S. Hahmann, and C. D. Hansen, Eds., 2003.

[11] H. Piringer, R. Kosara, and H. Hauser, "Interactive focus+context visualization with linked 2d/3d scatterplots," in *Proceedings of the Second International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'04)*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 49–60.

[12] R. Kosara, G. N. Sahling, and H. Hauser, "Linking scientific and information visualization with interactive 3d scatterplots," in *Short Communication Papers Proceedings of the 12th International Conference in Central Europe on Computer Graphics, Visualization, and Computer Vision (WSCG)*, 2004, pp. 133–140.

[13] N. Shah, V. Filkov, B. B. Hamann, and K. I. Joy, "Genebox: Interactive visualization of microarray data sets," in *Proceedings of The 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '03)*, F. Valafar and H. Valafar, Eds. Computer Science Research, Education, and Applications Press (CSREA), 2003, pp. 10–16.

[14] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics, Principles and Practice*, 2nd ed. Addison-Wesley, 1997, ch. 19.5.2.

[15] M. Z. Ludwig, N. H. Patel, and M. Kreitman, "Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change," *Development*, vol. 125, no. 5, pp. 949–958, 1998.